Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hybridization between Grid Search and Support Vector Machines for Predicting Stock Trend

Dr/ Abeer Badr El Din*
Lecture of Department Computer Science &Sadat Academy

Basant Ali Sayed Ali
Teaching, Dept. of Management Information System & Higher Institute of Qualitative Studies

Abstract: This paper presents a novel method for predicting stock trend by using Hybridization between Grid Search , pattern search and Support Vector Machines. The main aim of this paper displays how this Hybridization between Grid Search and Support Vector Machines can avoid the problem of overfitting .the paper introduce experiments performed on famous bench mark problem as empirical evidence .in addition it is applied on the case study of Egyptian stock market. The proposed approach has shown the efficiency and accuracy in the predicting the stock trend compared the classical support vector machines.

Key words: Stock Trend, support vector machines, grid search, pattern search, and Overfitting.

1- INTRODUCTION

The prediction of stock trend is too hard process in the field of finance. This area attracts many researchers to do their best in founding an accurate model. This area is complicated, uncertain, and has many uncontrollable variables. the fundamental approach has suffer from the practical difficulties in the real cases. In the same direction the technical analysis but really has better efficiency than fundamental approach. From this view the artificial intelligence be the panacea for predicting the stock trend. Among the approaches used in the artificial intelligence the neural network, the decision tree, and the modern trend is the support vector machines. SVM has many practical difficulties. The optimizing the parameter of machine learning is the core heart of process. This paper introduce a novel approach of the novel method for predicting stock trend by using Hybridization between Grid Search , pattern search and Support Vector Machines. The main aim of this paper displays how this Hybridization between Grid Search and Support Vector Machines can avoid the problem of overfitting .the remaining of the paper is structured as follow section two displays Support Vector Machines for Classification with More Than Two Categories, section three presents Optimal Fitting without over fitting. Section four introduces the proposed technique. Section five present the experiments, section six display conclusions.

2- SUPPORT VECTOR MACHINES FOR CLASSIFICATION WITH MORE MULTIPLE CATEGORIES

In the case of classification of multiple categories the core idea of availability of using hyper plane to realize the separation of the feature of vectors in two classes which performs related to the case of there are two categories, on the other hand new advanced research papers pay their attention to tackle the problem of how does SVM tackle the problem when the case has the target variable that has more than two categories? It is noticed the current researches has been motivated to find Several approaches, but no one can find more than the two common approaches that are the most popular: (1) the first approach is called one against many in this case it is found each category is split out and all of the remaining categories are merged; and, (2) the second approach where one against one where k(k-1)/2 models are established such k is the number of categories. The proposed technique uses the more accurate (but on the other hand more computationally cost) the second approach is technique of one against one. For a more details the discussion of why this working method is used and proved comparisons with other approaches see [Hsu and Lin, 2002].

3- OPTIMAL FITTING WITHOUT OVER FITTING

There is a serious problem in the field of the artificial intelligence especially in the support vector machines which is called overfitting problem .it is clear that the accuracy of the SVM technique is directly dependent on mechanism of how to select the kernel parameters for example C, Gamma, P, etc. the proposed technique provides two methods for trying to find optimal parameter values,

There are two methods of the parameter optimization is performed a grid search and a pattern search. The first method is the grid search. The main idea of these search methods tries values of each parameter across the certain search range based on the geometric steps for finding the optimal structure. The second approach is called the pattern search which also be known as a search of compass in some research papers is called a the method named search line) begins from centre of range of research and perform a series of steps For each direction for each parameter. But on the other hand, the case of fitting process exist of the model enhance, the centre of search transfer to the next point and the repetition of the process is continuous. But on the other hand, in the case of no improvement is exist, the step size is decay and this type of search can be repeated again. The pattern search can stop in the case of the search when step size is decay into a certain error.

In this paper to can avoid over fitting, the method of cross-validation is proposed to evaluate the model be fitting obtained by each value parameter set that try during search process of the grid or pattern The below figure introduced by Florian Markowitz illustrates how parameter be different values leads under or over fitting

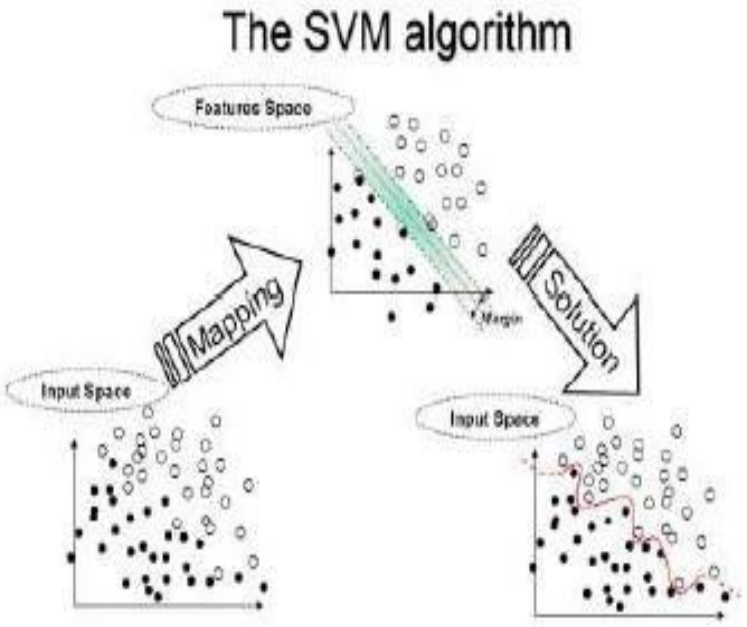

Fig1: illustrated the mechanism of SVM algorithm

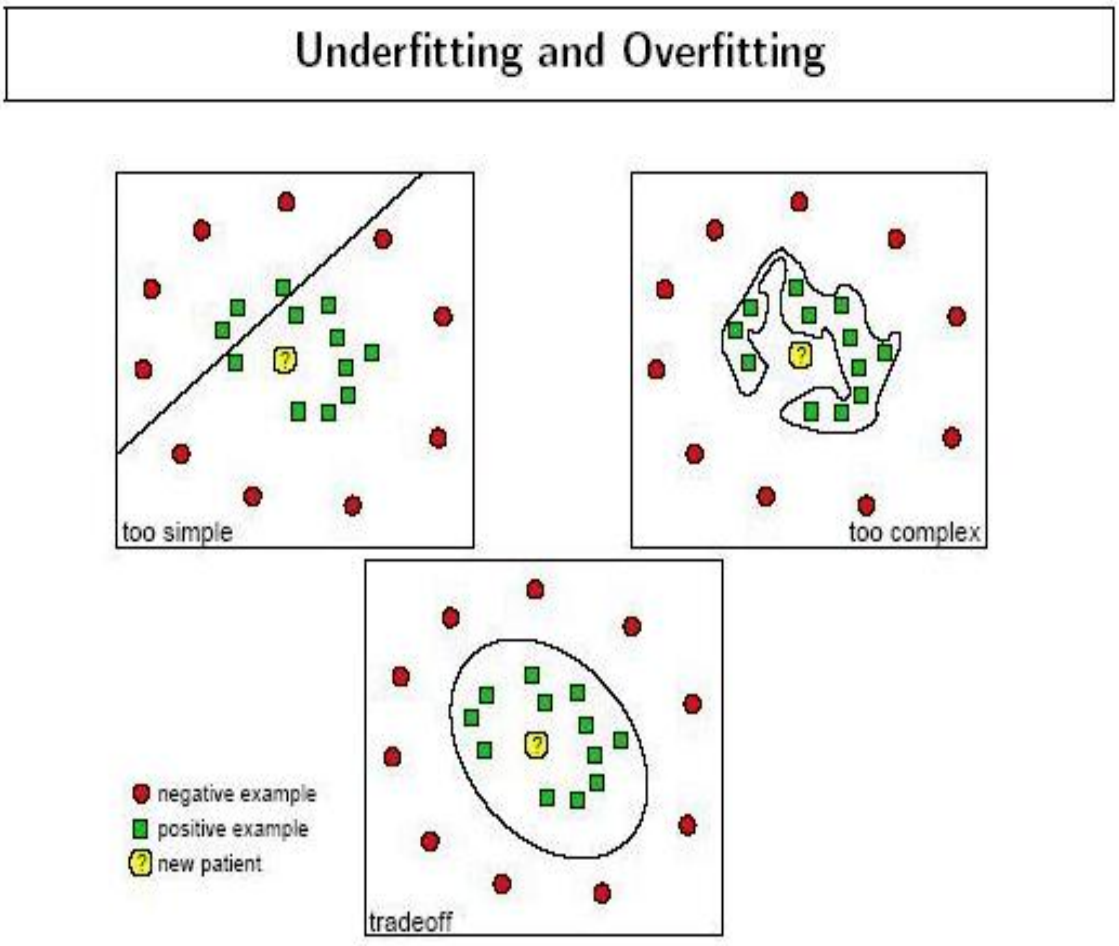

Fig 2: illustrated the under fitting and over fitting of SVM algorithm

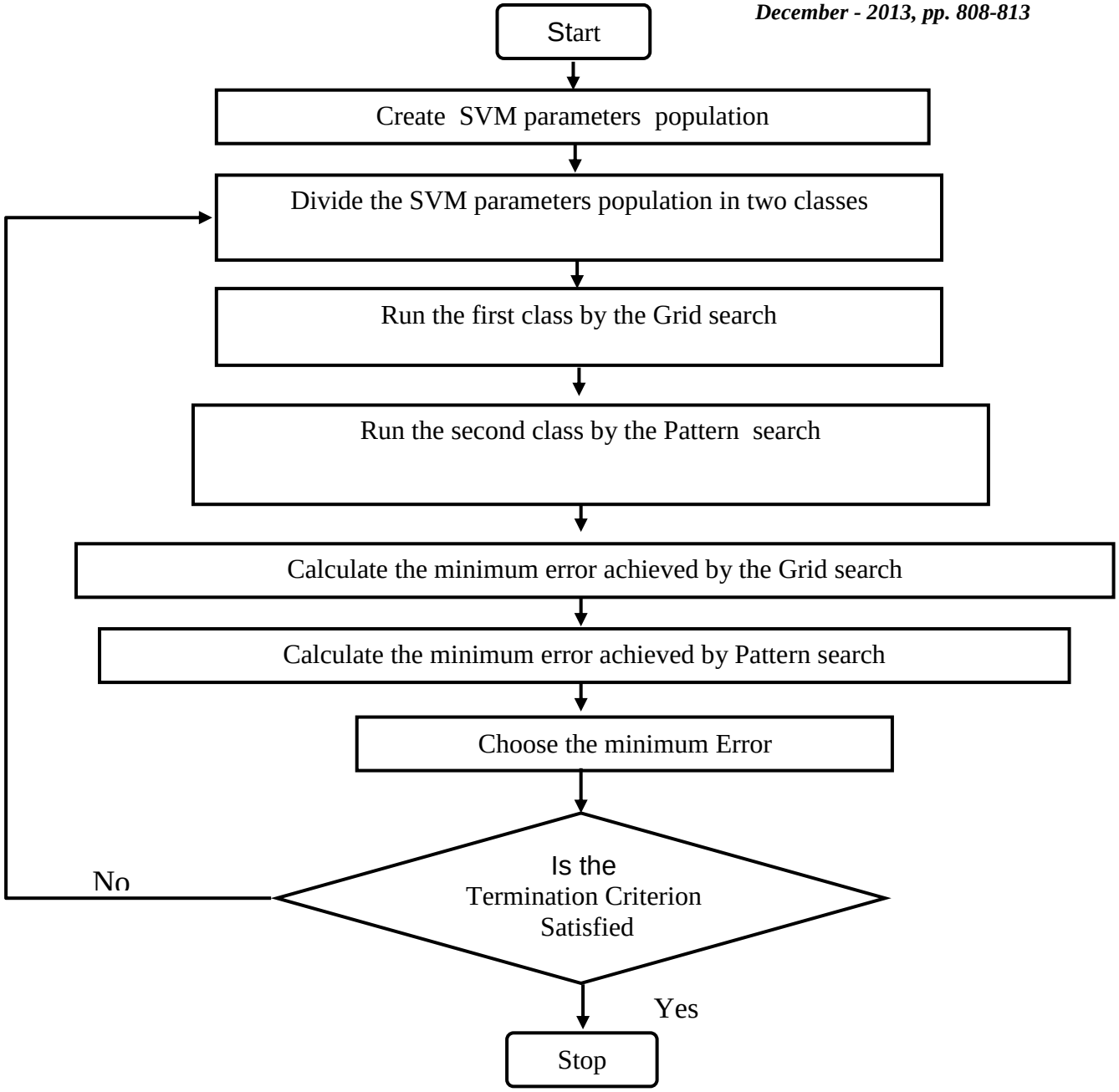

Fig3: flowchart of optimizing support vector machines to avoid overfitting

4- EXPERIMENTS AND RESULTS

4.1-Outliers

Outliers have bad effects on accuracy, particularly in the case of Grid search is implemented. The idea of the search is based on outlier of data pattern that can deviate substantially from the data distribution. Related to the huge deviation from the norm, outliers are lead to large errors, and consequently large weight updates. In previous work [Abeer Badr ELdin, Basant Ali Sayed, 2013]

The outlier problem can be addressed in the following ways: •

the first step of this process is Removing outliers before training starts based on statistical techniques. the software used in this work is Minitab software to detected the outliers and build the interquartile range according to the data of the case the results achieved as the following .

TABLE 1: DATA SUMMARY

Bootstrap Specifications

Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile

TABLE 2:DESCRIPTIVE STATISTICS

		Statistic	Std. Error	Bootstrap[a]			
				Bias	Std. Error	95% Confidence Interval	
						Lower	Upper
Open	N	799		0	0	799	799
	Range	15.48					
	Minimum	2.02					
	Maximum	17.50					
	Mean	6.4796	.10304	.0048	.1047	6.2956	6.6943
	Std. Deviation	2.91252		-.00572	.10544	2.70805	3.12495
	Variance	8.483		-.022	.615	7.334	9.765
	Skewness	1.262	.086	-.007	.084	1.085	1.422
	Kurtosis	1.993	.173	-.016	.326	1.372	2.662
Close	N	799		0	0	799	799
	Range	15.22					
	Minimum	2.02					
	Maximum	17.24					
	Mean	6.4641	.10250	.0045	.1043	6.2811	6.6780
	Std. Deviation	2.89746		-.00557	.10611	2.68781	3.10998
	Variance	8.395		-.021	.616	7.224	9.672
	Skewness	1.264	.086	-.006	.081	1.088	1.421
	Kurtosis	2.013	.173	-.011	.316	1.419	2.694
Low	N	799		0	0	799	799
	Range	14.74					
	Minimum	1.96					
	Maximum	16.70					
	Mean	6.2552	.09782	.0040	.0998	6.0765	6.4631
	Std. Deviation	2.76497		-.00623	.10107	2.56895	2.97107
	Variance	7.645		-.024	.560	6.600	8.827
	Skewness	1.233	.086	-.007	.082	1.058	1.389
	Kurtosis	1.974	.173	-.013	.313	1.391	2.631
High	N	799		0	0	799	799
	Range	15.82					
	Minimum	2.06					
	Maximum	17.88					
	Mean	6.6838	.10732	.0049	.1090	6.4913	6.9019
	Std. Deviation	3.03353		-.00537	.11074	2.81931	3.25232
	Variance	9.202		-.020	.673	7.949	10.578
	Skewness	1.287	.086	-.006	.082	1.117	1.447
	Kurtosis	2.040	.173	-.010	.320	1.423	2.730
volume	N	799		0	0	799	799
	Range	19900273.00					
	Minimum	52175.00					
	Maximum	19952448.00					
	Mean	2036678.5632	81904.36935	-2056.1420	79839.4092	1881270.2779	2189534.4238
	Std. Deviation	2315157.06789		-9312.95177	138407.53282	2034931.53574	2578201.47984
	Variance	5359952248985.000		-23897672649.191	639225569205.514	4140946792163.710	6647122875883.220
	Skewness	2.492	.086	-.058	.302	1.892	3.047
Valid N (listwise)	N	799		0	0	799	799

TABLE 3 : CASE PROCESSING SUMMARY

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Open	799	100.0%	0	0.0%	799	100.0%
Close	799	100.0%	0	0.0%	799	100.0%
High	799	100.0%	0	0.0%	799	100.0%
Low	799	100.0%	0	0.0%	799	100.0%
Volume	799	100.0%	0	0.0%	799	100.0%

SVM Misclassification Tables

TABLE 4: TRAINIG MISCLASSIFICATION

Actual			Misclassified			
Class	Actual count	Weight	Actual count	Weight	Percent of error	Error Cost
Down	417	417	123	123	**29.4**	0.294
Stable	30	30	30	30	**100.000**	1.000
UP	350	350	170	170	**48.5**	0.485
Total	797	797	323	195	**24.467**	0.245

Overall accuracy = 59.5%

TABLE 5: VALIDATION MISCLASSIFICATION TABLES

Actual			Misclassified			
Class	Actual Count	Weight	Actual Count	Weight	Percent of error	Error Cost
Down	417	417	200	200	**47.9**	0.479
Stable	30	30	30	30	**100.000**	1.000
UP	351	351	135	135	**38.462**	0.385
Total	798	798	365	245	**30.702**	0.307

Overall accuracy = 54.3%

Confusion Matrix for the Training stage

TABLE 6: TRAINING CONFUSION MATRIX

Actual	Predicted Category		
Category	Down	Stable	Up
Down	294	0	123
Stable	21	0	9
Up	114	0	236

Confusion Matrix for the Validation Stage

TABLE 7: VALIDATION CONFUSION MATRIX

Actual	Predicted Category		
Category	Down	Stable	Up
Down	217	0	200
Stable	21	0	9
Up	135	0	216

Hybridization between SVM and grid search Misclassification Tables

Misclassification for the Training Stage

TABLE 8: TRAINING MISCLASSIFICATION TABLES

Actual			Misclassified			
Class	Actual count	Weight	Actual count	Weight	Percent of error	Error Cost
Down	417	417	51	51	12.230	0.122
Stable	30	30	30	30	100.000	1.000
UP	350	350	114	114	32.571	0.326
Total	797	797	195	195	24.467	0.245

Overall accuracy = 75.53%

Misclassification For The Validation Stage

TABLE 9: VALIDATION MISCLASSIFICATION TABLES

Actual			Misclassified			
Class	Actual count	Weight	Actual count	Weight	Percent of error	Error Cost
Down	417	417	80	80	19.185	0.192
Stable	30	30	30	30	100.000	1.000
UP	351	351	135	135	38.462	0.385
Total	798	798	245	245	30.702	0.307

Overall accuracy = 69.30%

Confusion Matrix for the Training Stage

TABLE 10: TRAINING CONFUSION MATRIX

Actual	Predicted Category		
Category	Down	Stable	Up
Down	366	0	51
Stable	21	0	9
Up	114	0	236

Confusion Matrix for the Validation Stage

TABLE 11: VALIDATION CONFUSION MATRIX

Actual	Predicted Category		
Category	Down	Stable	Up
Down	337	0	80
Stable	21	0	9
Up	135	0	216

5- CONCLUSION

The results of experiments show the superior performance of the hybridization of grid search and support vector machines compared with the SVM without hybridization of grid search. The propsed technique proved its efficiency in avoiding the over fitting problem .this problem which is considered one of the difficult problem in the field of business.

6-REFERENCES

[1] Huber, P. (1964). Robust estimation of a location parameter, Annals of Math. Stat. 53: 73-101.
[2] Hsu, C.-W and C.-J. Lin. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 13(2):415-425, 2002.
[3] Huberty, Carl J. Applied Discriminant Analysis. John Wiley & Sons, 1994.
[4] Ivakhnenko G.A. Self-Organisation of Neuronet with Active Neurons for Effects of Nuclear Tests Explosions Forecasting. System Analysis Modeling Simulation.
[5] Julian, Randy. Using LDA. Lilly Research Laboratories (http://miner.chem.purdue.edu/Lectures/Lecture10.pdf).
[6] Klecka, William R. Discriminant Analysis. Sage Publications, 1980
[7] Kecman, Vojislav. Support Vector Machines Basics. School of Engineering Report 616. The University of Auckland, School of Engineering. April, 2004.
[8] Kleinbaum, David G., Mitchel Klein. Logistic Regression, A Self Learning Text, Second Edition. Springer, 1992.
[9] Kordík, Pavel, Pavel Náplava, Miroslav Šnorek, Marko Genyk-Berezovskyj. The Modified GMDH Method Applied to Model Complex Systems, Department of Computer Science and Engineering, CTU, FEE Karlovo nám. 13, Prague, Czech Republic
[10] Kubat, Miroslav and Stan Matwin. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection.
[11] Loh, W.Y. and Shih, Y.S. (1997). Split selection methods for classification trees. Statistica Sinica 7: 815-840.
[12] Maindonald, John and John Braun. Data Analysis *and Graphics Using R, An Example-based Approach*. Cambridge University Press, 2003.
[13] Markowetz florian Classification by Support Vector Machines Practical DNA Microsarray Analysis 2003 Max Planck Institute for Moleculer Genetics Computational Moleculer Bioligy, Berlin, https://phssec1-fhcrc.org/secureplone/www.bioconductor.org/workshops/2003/NGFN03/svm.pdf
[14] Abeer Badr El Din ,Basant Ali. Predicting the Stock Trend by Artificial Intelligence Techniques based on Structural Risk Minimization,2013.

